2019-08-14

An Experimental Validation of Masking in IEC 60601-1-8:2006-compliant Alarm Sounds

Bolton, M

10.1177/0018720819862911
Human Factors: the journal of the human factors and ergonomics societ
SAGE Publications

Health Care/Health Systems

An Experimental Validation of Masking in IEC 60601-1-8:2006-compliant Alarm Sounds

Matthew L. Bolton, Xi Zheng, Meng Li
University at Buffalo, the State University of New York, Buffalo, NY, USA

Judy Reed Edworthy
University of Plymouth, Plymouth, England, UK
Andrew D. Boyd

University of Illinois at Chicago, Chicago, IL, USA

Author Note

Manuscript Type: Extended Multi-Phase Study

Word Count: 7,081

The research reported in this paper was supported by the Agency for Healthcare Research and Quality under award number R18HS024679. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University at Buffalo. Informed consent was obtained from each participant.

Correspondence concerning this article should be addressed to Matthew L. Bolton, Department of Industrial and Systems Engineering, University at Buffalo, the State University of New York, Buffalo, NY 14260. Contact: mbolton@buffalo.edu

Abstract

Objective: This research investigated whether the psychoacoustics of simultaneous masking, which are integral to a model-checking-based method, previously developed for detecting perceivability problems in alarm configurations, could predict when IEC 60601-1-8-compliant medical alarm sounds are audible. **Background:** The tonal nature of sounds prescribed by IEC 60601-1-8 makes them potentially susceptible to simultaneous masking: where concurrent sounds render one or more inaudible due to human sensory limitations. No work has experimentally assessed whether the psychoacoustics of simultaneous masking accurately predict IEC 60601-1-8 alarm perceivability. **Methods:** In two signal detection experiments, 28 nursing students judged whether alarm sounds were present in collections of concurrently sounding standard-compliant tones. The first experiment used alarm sounds with single-frequency (primary harmonic) tones. The second experiment's sounds included the additional, standard-required frequencies (often called sub-harmonics). T-tests compared miss, false alarm, sensitivity, and bias measures between masking and non-masking conditions and between the two experiments. **Results:** Miss rates were significantly higher and sensitivity was significantly lower for the masking condition than for the non-masking one. There were no significant differences between the measures of the two experiments. **Conclusion:** These results validate the predictions of the psychoacoustics of simultaneous masking for medical alarms and the masking detection capabilities of our method that relies on them. The results also show that masking of an alarm's primary harmonic is sufficient to make an alarm sound indistinguishable. **Application:** Findings have profound implications for medical alarm design, the international standard, and masking detection methods.

Keywords: Medical devices and technologies, Audition, Patient safety, Psychophysical methods, Signal detection theory

Précis: This research used signal detection experiments to validate that the psychoacoustics of simultaneous masking could predict when IEC 60601-1-8-compliant medical alarm sounds are audible based on the masking of alarm primary harmonics. The results will inform the IEC 60601-1-8 standard and methods for detecting masking in medical alarms.

An Experimental Validation of Masking in IEC 60601-1-8:2006-compliant Alarm Sounds

Introduction

In modern medical environments, a single patient produces hundreds of alarm per day and thus tens of thousands of alarms are generated a day in a given hospital (The Joint Commission, 2013a). Health professionals are not always responding to these alarms, and this is a very dangerous problem. The Pennsylvania Patient Safety Authority (2009) reported 194 problems (12 that resulted in fatalities) with medical personnel failing to react to telemetry monitoring alarms from June 2004 through December 2008. Similarly, 98 alarm nonresponses (5 extended patient hospital stays, 13 produced "permanent loss of function," and 8 ended in patient death) were documented in a Sentinel Event Alert (The Joint Commission, 2013a) that covered a period from January 2009 to June 2012. Because of these types of problems, the ECRI Institute has consistently named medical alarms one of the most important technological hazards to patient safety for more than a decade (ECRI Institute, 2018; Stead & Lin, 2009).

There are many reasons why humans may fail to respond to medical alarms including the number of false alarms, the lack of consistent design philosophies between alarms and medical devices, and designs that do not facilitate alarm learnability and discernibility (see reviews by Edworthy 2013; Edworthy et al. 2018). The perceivability of the alarms in the presence of other alarms is at least partially responsible for this problem (ECRI Institute, 2014; The Joint Commission, 2013a, 2013b; Vockley, 2014).

One issue that can impact the perceivability of medical alarms is simultaneous masking. In simultaneous masking, limits of the human sensory system prevent humans from hearing one or more concurrent sounds (Fastl & Zwicker, 2006). A number of researchers have generally speculated that simultaneous masking could be a problem with medical alarms (Edworthy & Hellier, 2005, 2006; Edworthy & Meredith, 1994; Konkani, Oakley, & Bauld, 2012; Meredith & Edworthy, 1995; Patterson, 1982; Patterson & Mayfield, 1990). This is because many medical alarms are often represented as melodies of tonal sounds, including alarms that are compatible with the International Electrotechnical Commission's (IEC's) international standard (IEC 60601-1-8:2006+AMD1:2012). This makes them especially prone to simultaneous masking (Bosi & Goldberg, 2003; Fastl & Zwicker, 2006). There is also empirical evidence that simultaneous masking does occur for medical alarms in modern hospitals. Momtahan, Hetu, and Tansley (1993) analyzed 49 different medical alarms and found 25 pairs in which one could be completely masked by the other. Toor, Ryan, and Richard (2008) discovered several instances where high priority alarms could be masked by lower priority operating room sounds including other alarms, telephone rings, and beeper sounds. Both of these studies involved recording sounds in a medical environment and then using the psychoacoustics of simultaneous masking (mathematical formulations that predict whether simultaneous masking occurs based on the volumes and frequencies of the sounds; Bosi and Goldberg 2003) to identify pairs of alarms where masking could occur.

Despite these findings, medical alarm safety has mostly focused on other problems (Edworthy, 2013). This is likely due to the complexity of simultaneous masking. Masking can manifest as a result of multiple simultaneously sounding alarms (not just pairs) and may only

occur for particular timings of the overlaps between the alarms. It is thus almost impossible for analysts to experimentally determine how masking could manifest in alarm configurations. Given the sheer number of medical alarms and possible different overlaps between them in a given hospital (The Joint Commission, 2013a), it is likely that masking is an important factor in alarm nonresponse.

To address this situation, we developed a computational method (Bolton, Edworthy, & Boyd, 2018; Bolton, Edworthy, Boyd, Wei, & Zheng, 2018; Bolton, Hasanain, Boyd, & Edworthy, 2016; Hasanain, Boyd, & Bolton, 2016, 2014; Hasanain, Boyd, Edworthy, & Bolton, 2017) that uses the psychoacoustics of simultaneous masking and model checking. Model checking is a formal method that allows an analyst to automatically, mathematically prove properties against models of concurrent systems (a process called formal verification; Clarke, Grumberg, and Peled 1999). In our method, an analyst models the behavior of alarms and runs model checking to prove if the represented alarms can ever mask each other. This method has been used to analyze real medical alarm configurations (Bolton, Edworthy, Boyd, Wei, & Zheng, 2018; Bolton et al., 2016; Hasanain et al., 2017) and the reserved alarm sounds of the IEC 60601-1-8 international standard (Bolton, Edworthy, & Boyd, 2018).

This method is powerful and offers unprecedented masking detection capabilities. However, the method has limitations. First, like the experimental results presented by Momtahan et al. (1993) and Toor et al. (2008), the method relies on the psychoacoustics of simultaneous masking. While these psychoacoustics have been well tested over the years (Bosi & Goldberg, 2003), they have not been explicitly experimentally validated for medical alarm sounds. Second, many tonal medical alarms are consistent with the IEC 60601-1-8 standard. This means that they contain a primary harmonic (frequency) as well as several additional harmonics (usually the minimum of 4) that are multiples of the primary that are at lower volumes. While our method is capable of accounting for the masking effects of both primary and additional harmonics, including the additional harmonics can require orders of magnitude more computational time. Thus, if the masking of the primary harmonics was critical to alarm perceivability irrespective of the additional harmonics, this would profoundly improve the usefulness and relevance of our method.

We addressed both of these issues by conducting two signal detection theory (SDT) experiments. In the first, we validated the ability of our method to predict masking between primary harmonics of IEC 60601-1-8 medical alarm sounds. In the second, we assessed how well predictions about masking between the primary harmonics impact the perceivability of alarms with a full set of IEC 60601-1-8-required additional frequencies.

Review of Relevant Literature

Below we provide background on the alarms of IEC 60601-1-8, the psychoacoustics of simultaneous masking that are used by our method to predict masking, and the SDT experimental paradigm that we use in our research.

IEC 60601-1-8

The IEC 60601-1-8 international medical alarm standard is widely used across the medical industry. It was created to improve alarm discernibility and identification. As part of this, it provides instructions for designing new alarm sounds, which typically manifest as

melodies (sequences) of tones separated by pauses. There are many details in the standard. For the work presented in this paper, we are primarily concerned with the specific requirements of the individual tones that compose alarm melodies. Each tone in a melody has a single primary frequency. It also has several additional harmonics (additional frequencies) designed to make the alarms more tonally rich and help listeners localize alarm sources. The standard does not require specific frequencies, volumes, and timings of the tones in alarm melodies. Rather, it provides ranges of acceptable values. These are summarized in Table 1.

Table 1: IEC 60601-1-8 Alarm Tone Characteristics

Tone Characteristic	Value Range
Primary Frequency (Hz)	[150, 1000]
Primary Frequency Volume (dB)	$\boldsymbol{v}$
Maximum Primary Tone Volume Difference (dB)	10
Minimum Number of Additional Harmonics	4
Additional Harmonics Frequency (Hz)	[300, 4000]
Additional Harmonic Volume (dB)	[v-15, v+15]
Duration (s)	[0.075, 0.25]

A number of issues have been identified with the melodic alarm sounds prescribed in the standard that compromise the standard's goal of making alarms discernable and identifiable (Edworthy et al., 2018). In this work, we are particularly concerned with the effect simultaneous masking has on alarm audibility.

Masking and the Psychoacoustics of Simultaneous Masking

Auditory masking describes a number of different phenomena where a sound is rendered inaudible due to the presence of one or more other (masking) sounds. For example, pressure waves of sounds can physically interact to cancel each other out or a given sound can be indistinguishable from environmental noise. In this work, we focus on simultaneous masking. This occurs when similar, simultaneous sounds render one or more imperceptible due to the way that the sounds affect the sensitivity of the human sensory system.

Our method uses of the psychoacoustics of simultaneous masking to make predictions about whether any given alarm in a configuration will be audible. The psychoacoustics of simultaneous masking mathematically describe how the volumes and frequencies of sounds produce masking. In particular, the psychoacoustics are based on how masking sounds (*masker*s) stimulate the sensors of the basilar membrane: the spiral-shaped physical structure in the human inner ear that is responsible for the ability of humans to distinguish between sounds (Ambikairajah, Davis, & Wong, 1997; Baumgarte, Ferekidis, & Fuchs, 1995; Bosi & Goldberg, 2003; Brandenburg & Bosi, 1997; Brandenburg & Stoll, 1994; Schroeder, Atal, & Hall, 1979). This raises the absolute threshold (in dB) that the volume of another sound (a potential *maskee*) must exceed to be perceivable (Bosi & Goldberg, 2003).

These psychoacoustics render frequencies on the Bark scale (E. Zwicker and R. Feldtkeller, 1967): a scale that maps a frequency in Hz to a position on the basilar membrane where that frequency most powerfully stimulates the receptors (see Figure 1). A frequency in Hz (f_{sound}) is converted into Barks using Equation 1.

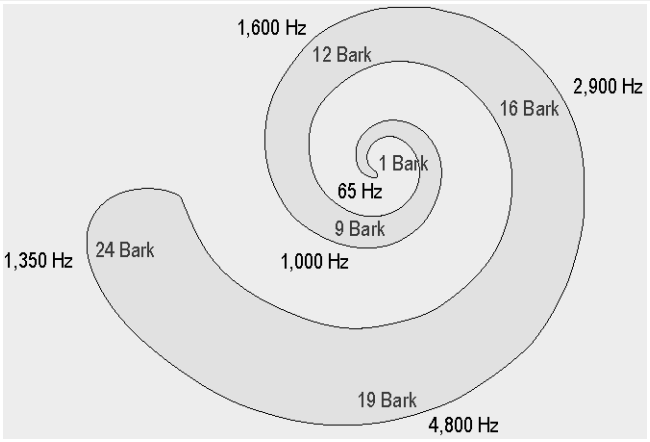

Figure 1: Depiction of how peak stimulation of sounds in Hz occurs at different Bark locations along the basilar membrane.

$$z_{sound} = 13 \cdot \arctan(76 \cdot f_{sound}/100000) + 3.5 \cdot \arctan\left((f_{sound}/7500)^2\right) \quad (1)$$

$$\text{curve}_{masker}(z_{maskee}) = \text{spread}_{masker}(\delta z) v_{masker} - \Delta \quad (2)$$

$$\delta z = z_{maskee} - z_{masker}. \quad (3)$$

$$\text{spread}_{masker}(\delta z) = \begin{cases} -17 \cdot \delta z + 0.15 \cdot v_{masker} \cdot (\delta z - 1) \cdot \theta(\delta z - 1) & \text{for } \delta z \geq 0 \\ -(6 + 0.4 \cdot v_{masker}) \cdot |\delta z| - (11 + 0.4 \cdot v_{masker} \cdot (|\delta z| - 1)) \cdot \theta(|\delta z| - 1) & \text{otherwise} \end{cases} \quad (4)$$

where $\theta(x) = 1$ for $x \geq 0$ and $\theta(x) = 0$ otherwise

$$\Delta = 6.025 + 0.275 \cdot z_{masker} \text{ dB}. \quad (5)$$

$$\text{power}(v) = 10^{v/10} \quad (6)$$

$$\text{power}(\text{mthresh}_{maskee}) = \text{power}(\text{abs}_{maskee}) + \left(\sum_{n=1}^{N} \text{power}(\text{curve}_{masker_n}(z_{maskee}))^{\alpha}\right)^{1/\alpha} \quad (7)$$

$$\text{abs}_{maskee} = 3.64 \cdot (f_{maskee}/1000)^{-0.8} - 6.5 \cdot e^{-0.6(f_{maskee}/1000 - 3.3)^2} + 10^{-3} \cdot (f_{maskee}/1000)^4 \quad (8)$$

The "masking curve" calculates how a given masker shifts the absolute threshold of hearing with Equation 2. In this, v_{masker} is the masker's volume in dB and δz is calculated using Equation 3, where z_{maskee} and z_{masker} are the Bark scale frequencies of the maskee and masker respectively. Further, the $spread_{masker}$ (Equation 2) function model show the magnitude/volume of the masking threshold changes with respect to δz. Finally, Δ is the minimum difference between the volumes of the masker and maskee that can result in masking.

There are multiple formulations of the psychoacoustic spreading function and Δ based on the characteristics of the masking and masker sounds. In this research, we use the formulation in Equation 2 for the masking curve and the Δ formulation in Equation 5. These were used because they are universally regarded as the most accurate for modeling tonal sounds (Ambikairajah et

al., 1997; Bosi & Goldberg, 2003; Brandenburg & Stoll, 1994). Figure 2 illustrates the shape of the masking curve described by Equation 2.

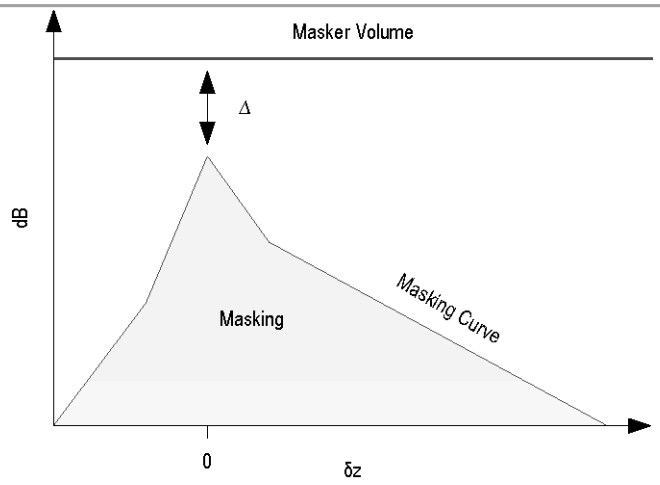

Figure 2: The masking curve shape dictated by Equations 2 to 5.

Furthermore, the combined masking threshold of multiple concurrent sounds can be greater than a simple sum of the effect of individual maskers (Bosi & Goldberg, 2003; Humes & Jesteadt, 1989). This effect is called additive masking. Because masking levels are measured in dB (a logarithmic scale), these are transformed to the power scale to allow for arithmetic operations. A volume in dB (v) can be converted to the power scale using Equation 6. Then, for a given potential maskee and N potential maskers, the absolute value of hearing adjusted for the additive effect of masking (in dB) is calculated using Equation 7. In Equation 7, $\boldsymbol{\alpha}$ is a positive constant (Green, 1967). abs_{maskee} is the unaltered absolute threshold of hearing (in dB) at the maskee's frequency. This, using the maskee's frequency in Hz, is described using Equation 8 (Terhardt, 1979).

These psychoacoustics have been used successfully to predict masking for normal human hearing for decades (Bosi & Goldberg, 2003). They were employed by researchers to identify when masking could occur for sounds recorded in medical environments (Momtahan et al., 1993; Toor et al., 2008). They were also the basis for lossy audio compression techniques (like those used in MPEG [Moving Picture Experts Group] formats; Bosi and Goldberg 2003), digital audio compression methods that allow reductions in the size of audio files by removing audio data that is predicted to be masked.

Signal Detection Theory

Signal detection theory models the detection of an event in a noisy environment. In a human judgment context, this captures both the state of the world (whether there is signal in the presence of noise or just noise) and the human's response ("Yes" there is a signal or "No" there is no signal). Based on this representation, there are four possible classifications of the outcome (Figure 3). Two of these are correct. If the judge says "Yes" when there is signal, the outcome is a hit. If the judge says "No" when there is only noise, the outcome is a correct rejection. Two of the outcomes are incorrect. If the judge says "Yes" when there is only noise, the outcome is a false alarm. If the judge says "No" when there is a signal, the outcome is a miss.

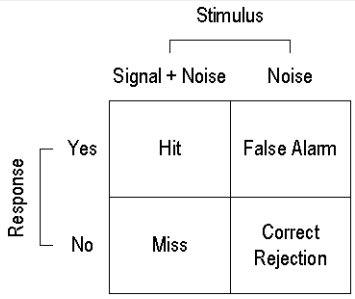

Figure 3: A matrix describing how the different outcomes can manifest based on a "Yes" or "No" human response to a stimulus that is either signal or noise.

When a human performs signal detection task and makes multiple judgments in response to different states of the world, rates can be calculated for each of the outcomes:

Hit Rate:

$$H = \frac{\#\text{ of hits}}{\#\text{ of signal events}} \tag{9}$$

Miss Rate:

$$M = \frac{\#\text{ of misses}}{\#\text{ of signal events}} = 1 - H \tag{10}$$

False Alarm Rate:

$$F = \frac{\#\text{ of false alarms}}{\#\text{ of noise events}} \tag{11}$$

Correct Rejection Rate:

$$C = \frac{\#\text{ of correct rejections}}{\#\text{ of noise events}} = 1 - F \tag{12}$$

Note that because of the inverse relationships between hits and misses and between false alarms and correct rejections, analysts will typically only discuss results from one rate from each pair. For example, in the presented work, we only talk about miss and false alarm rates.

Two additional measures for modeling human judgment are typically calculated from the above rates: sensitivity and response bias (or simply bias). Sensitivity captures the judge's ability to distinguish signal from noise. Response bias is a measure of whether a judge is more likely to respond one way or another.

When the signal and noise can be assumed to be normally distributed with equal variance, sensitivity is the distance between the means of the signal and the noise distributions. The response bias is the likelihood ratio that a response of "Yes" is due to the presence of signal as opposed to noise alone. However, for many judgment tasks, the distributions of signal and noise may not be normally distributed (as will be the case in the experiments presented in this paper) or the distributions may be unknown. Thus, there are non-parametric measures for

computing sensitivity and response bias (Macmillan & Creelman, 1990; See, Warm, Dember, & Howe, 1997). In this work, we use the nonparametric calculations that have been shown to be appropriate in human subject experiments (See et al., 1997).

A', based on concepts introduced by Pollack and Norman (1964), calculates nonparametric sensitivity by approximating the area under a receiver operating characteristic (ROC) curve defined by the observed hit (H; Equation 9) and false alarm (F ; Equation 11) rates (Snodgrass & Corwin, 1988):

$$A' = \begin{cases} 0.5 + \frac{(H-F)\cdot(1+H-F)}{4\cdot H\cdot(1-F)} & \text{if } H \geq F \\ 0.5 + \frac{(F-H)\cdot(1+F-H)}{4\cdot F\cdot(1-H)} & \text{otherwise.} \end{cases} \tag{13}$$

This produces a value between 0 and 1, where a higher value indicates that the judge was more sensitive (more readily able to distinguish between signal and noise).

B''_D, which was introduced by Donaldson (1992), is a nonparametric measure of response bias that is also based on the geometry of the ROC curve:

$$B''_D = \frac{(1-H)\cdot(1-F) - H\cdot F}{(1-H)\cdot(1-F) + H\cdot F}. \tag{14}$$

A $\boldsymbol{B''_D}$ bias will range between -1 and 1, where a negative value indicates that the judge is more likely to say no (has a conservative bias), a positive value indicates that the judge is more likely to say yes (has a liberal bias), and a value of zero indicates that the judge is just as likely to say one or the other.

Experiment 1

In our first experiment, we used a SDT procedure to assess how well the psychoacoustics of simultaneous masking that are used in our method predict the ability of humans to perceive the primary harmonics of alarms sounds from IEC 60601-1-8.

Methods

Participants

A power analysis revealed that 80% power was achieved for detecting a moderate effect size ($d = 0.55$) with a two-tailed paired t-test with 28 participants. Thus, 28 participants were recruited for this study. Nursing students from the University at Buffalo were used as the participant pool because it constituted members of the actual population that will experience medical alarm sounds in a natural environment. Twenty one of the recruited students were female and seven were male. The experiment did not control for musical experience because we could find no research showing that musical ability had any impact on masking.

Materials and Apparatus

The experiment was run in the Usability Laboratory at the University at Buffalo, a controlled, quiet, evenly-lit environment. It was administered on a laptop computer resting on a computer desk (see Figure 4) in front of which a participant would sit. The laptop computer was connected to an external USB, 7.1 sound card. Four single-driver computer speakers were connected to the sound card so that each speaker only output sounds sent to a single channel of the sound card. The speakers were placed in line with each other on an elevated platform behind the laptop. The laptop computer was also connected to an optical computer mouse that the participant could use to interact with the software that administered the experiment.

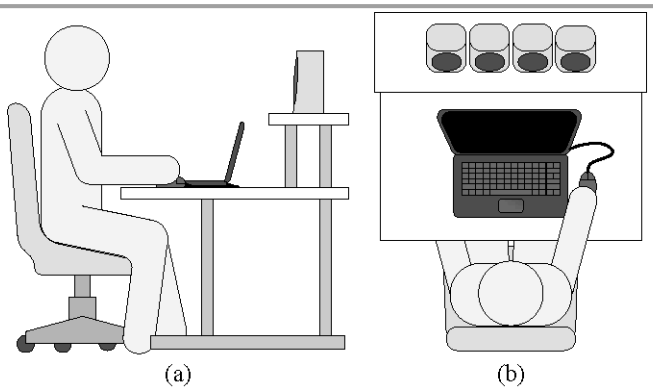

Figure 4: The physical apparatus setup used in the reported experiments. Both depict (in profile (a) and from above (b)) a participant sitting in front of a computer desk on which a laptop computer, a computer mouse, and four speakers were placed.

The software used for administering the experiment was constructed specifically for this study. This was implemented as a Visual Basic for Applications program within a Microsoft Excel spreadsheet. This software was able to examine the experimental design (which was stored in the spreadsheet), administer a given participant's experiment according to it, collect user responses, and store them in a separate excel sheet. The interface that the software used for administering the experiment is shown in Figure 5. This told a participant which trial they were on, out of the total number of trials. It also gave participants instructions for how to perform the trial.

Figure 5: The interface to the software used to administer the experiment and collect participant responses. This was always displayed in full screen so that the user could not see or interact with the Excel spreadsheet running in the background.

In a given trial, participants were charged with determining whether a judgment sound was present in a test sound. The judgment sound represented a single alarm sound that was always played on the rightmost speaker (Figure 4). The test sound constituted a simulation of the simultaneous sounding of multiple alarm sounds (between one and three) from different devices. Thus the alarm sounds of the test sound were each played on one of the three left-most speakers (one sound per speaker; Figure 4). When interacting with a trial in our software, a participant would first click on the "Play Judgment Sound" button to play the judgment sound. Participants would then click on the "Play Test Sound" button to play the test sound (Figure 4). Participants were allowed to play either of these sounds (one at a time) as many times as they wanted to until they felt like they could render a judgment. When participants were ready, they would indicate whether or not they thought the judgment sound was present in the test sound by clicking on the "Yes" (indicating they they thought the sound was present) or "No" (indicating that they thought the sound was not present) radio buttons. When participants were satisfied with their answers, they would click on the "Next" button. The interface would then present a dialogue box that would ask participants if they wanted to confirm their answer. If participants pressed a "No" button, they would stay on the current trial. If they pressed a "Yes" button, they would go to the next trial. Whenever this dialog box was being displayed, the software played brown noise (signal noise naturally produced by Brownian motion; Vasseur and Yodzis 2004) from the speakers to give participants a "palate cleanser" between trials.

A sound level meter, positioned at the ear position of a participant, was used to calibrate the laptop and speakers so that volumes matched the levels specified by the experiment.

A given trial was a pair of sounds: the judgment sound (a single alarm sound played on a single speaker in the apparatus) and the test sound (a collection of one to three alarm sounds, each played synchronously on a separate speaker in the apparatus). All of the sounds were designed to be consistent with the requirements of single tones from IEC 60601-1-8-compliant alarm sounds (Table 1), with only the primary harmonics. Each of the sounds was 0.25 seconds long. The tone in the judgment sound was 70 dB. One of the tones in all of the test sounds was 70 dB. The other tones in the test sound were 85 dB. If the judgment sound was in the test sound, the judgment sound was always the 70 dB sound in the test sound. These volumes were used because they were allowable by the standard (which specifies variations in volumes of any alarm sounds at the same priority be within 15 dB of each other; Table 1), are consistent with alarm volumes used in the field, and were not loud enough to cause hearing problems when combined together in the experiment. The set of frequencies used for tones were based on piano notes that fit within the allowable range of the standard (and are used to formulate the reserved alarm sounds in the standard; see Table 2). The frequencies used for the tones of the test sound were always different from each other.

Table 2: The frequencies used in tones found in judgment and test sounds.

Scientific Pitch Notation	Frequency (Hz)
C_4	261.63
$C\sharp_4$	277.18
D_4	293.66
$D\sharp_4$	311.13
E_4	329.63
F_4	349.23
$F\sharp_4$	369.99
G_4	392.00
$G\sharp_4$	415.30
A_4	440.00
$A\sharp_4$	466.16
B_4	493.88
C_5	523.25

Independent Variables

There were two independent within-subject variables in the experiment that enabled the use of a SDT experimental design. First, a trial was either a signal trial or a noise trial. In a signal trial, the judgment sound was one of the sounds output in the test sound. In a noise trial, the judgment sound was not part of the test sound. Second, a trial could either contain masking (where the 70 dB tone was masked by the other tones in the test sound according to the psychoacoustics of simultaneous masking) or not (where, according to the psychoacoustics, none of the tones in the test sound would be masked). In trials that were both signal and masking, the

test sound would contain the judgment sound and the judgment sound would be the one predicted to be masked.

Dependent Measures

For each trial, participants would indicate whether they thought the judgment sound was present in the test sound. This "Yes" (the judgment sound was present) or "No" (the judgment sound was not) response was the only dependent measure in the experiment.

Procedure

In the experiment, a participant was admitted to the lab and sat in front of the apparatus as shown in Figure 4. The participant was then given an informed consent document which they read and signed. After this, participants were read instructions that told them how to interact with the software interface (Figure 5) to administer the experiment. The participants were given a copy of the instructions for their reference. The participant then interacted with the software's interface to administer training and the experiment. When the experiment was completed, participants were given a $20 Amazon gift card.

Training

Before the proper start of the experiment, all participants experienced the same 18 training trials that were designed to introduce them to the judgment task. This was done by presenting trials in blocks. All trials and blocks were always presented to participants in the same order. The first block of four trials were signal trials that did not contain masking. The second block of four trials were noise trials that also did not contain masking. The third block of four trials were signal trials that did contain masking. For all three of these blocks, dialog boxes introduced the blocks and told patients whether he or she should or should not hear the judgment sound in the test sound. In the final block of six trials, the trials were a random ordering of: two signal trials with masking, two signal trials without masking, one noise trial without masking, and one signal trial without masking. In this final block, participants were told it was up to them to determine if the test sound was in the judgment sound. Across all of the training trials, participants were given feedback, via a dialog box, about the accuracy of each judgment after it was made.

Experimental Design

Following training, each participant experienced the same 200 experimental trials. These trials were grouped in a single block and were arranged consistently with the standards for non-parametric, human subjects, SDT designs as outlined by McNicol (2005), who recommended 50 masking and 50 noise trials for each experimental condition considered in an experiment. As per these standards, trials contained 100 masking trials and 100 trials without masking, where there were 50 signal and 50 noise trials in each 100 trial designation. All 200 trials were presented to each participant in a unique, randomly generated order. In signal trials, the speaker on which the judgment sound was played as part of the test sound was counterbalanced between trials.

The number of tones included in the trial's test sound could vary. In masking trials, test sounds could have either two and three tones (there were equal numbers of masking trials with

each number of tones). In trials without masking, test sounds could have between one and three tones (there were equal numbers of non-masking trials with each number of tones). Test sounds in masking trials were not allowed to have one tone because simultaneous masking could not occur in such a situation. Test sounds were allowed to have one tone in trials without masking because it could provide a non-masking condition that was perceptually comparable to a masking condition with two tones.

Data Analysis

Because simultaneous masking theoretically makes masked alarms inaudible, we hypothesized that we would observe a significantly higher miss rate (*M*) for the masked trials than the unmasked ones. Due to the nature of SDT rates (Equations 9 to 12), this would correspond to a significantly lower hit rate (*H*) for masked trials than the unmasked trials. Because the presence or absence of masking should not impact a human's tendency to say "Yes" in noise trials, we did not hypothesize a significant difference in false alarm rates (*F*) (and thus correct rejection rates; *C*) between masked and unmasked trials.

Because the inability to hear alarms would suggest a drop in human sensitivity, we hypothesized that humans would exhibit a lower sensitivity that was significant for masked trials than for unmasked ones. We did not hypothesize that the presence of masking would impact participant response bias.

To test these hypotheses, we analyzed each participant's responses in accordance with the SDT measures discussed in the background section. First, for each participant, the masking and non-masking trials were analyzed separately and used to compute each of the SDT rates (*H*, *F*, *M*, and *C*; Equations 9 to 12) and their associated nonparametric measures of sensitivity (A'; Equation 13) and bias (B''_D; Equation 14). Then, we used paired t-tests to compare *M*, *F*, A', and B''_D across participants. For *M* and A', because we hypothesized a direction to differences, one-tailed tests were used. For the other measure, because no direction of difference was hypothesized, two-tailed tests were used. Ultimately, statistical significance was assessed at an alpha level of 0.05 that was Bonferroni adjusted for the 10 different t-tests performed for the research presented in this paper. This ultimately resulted in an adjusted significance level of $0.05/10 = 0.005$. Effect sizes of these tests were computed using a Cohen's *d*.

Note that to ensure that the assumptions for the t-tests were valid, in all cases, an Anderson Darling test was conducted to assess the normality of the difference between the paired rates of participants.

Results

The results of the comparisons of miss and false alarm rates (*M* and *F* respectively) are reported in Figure 6. These analyses showed that miss rate (*M*) was significantly higher for masking trials than for trials without masking. There was no significant difference between false alarm rates (*F*) between masking and non-masking trials.

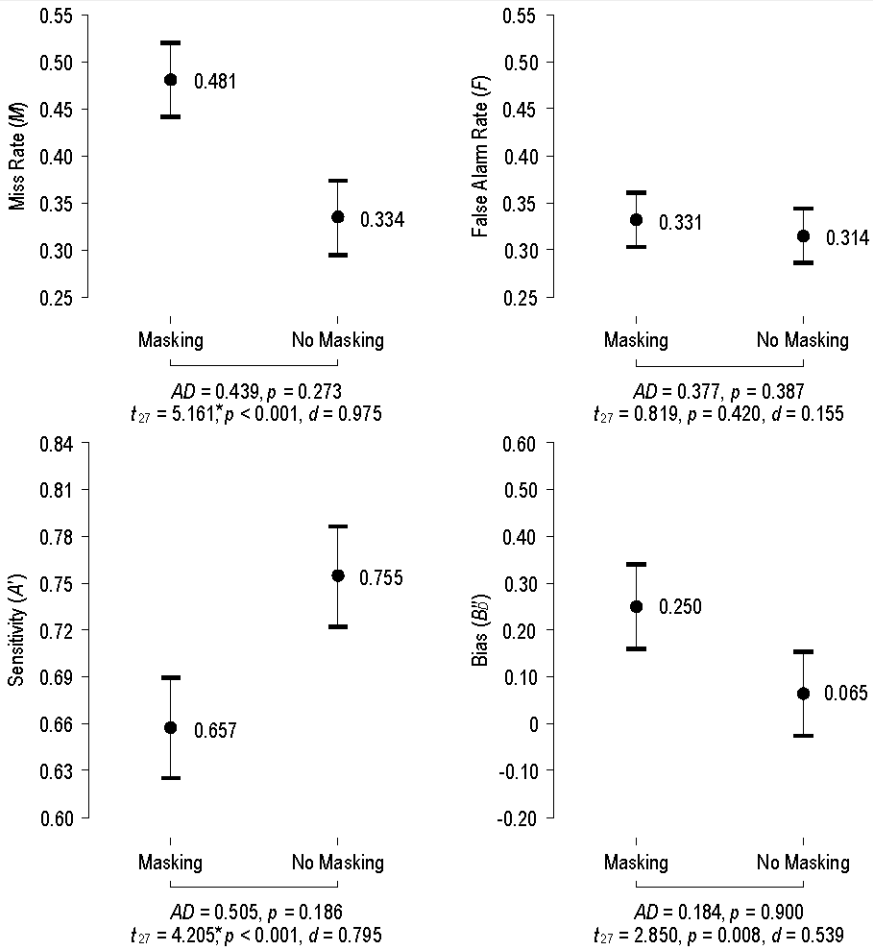

Figure 6: Means (labeled circles) and 95% within-subject confidence intervals (horizontal bars; Cousineau and O'Brien 2014) for miss rates (*M*), false alarm rates (*F*), sensitivity (*A'*) and bias (B''_D) for both the masking and non-masking conditions observed during experiment 1. Rates are presented with Anderson Darling statistics that indicate that differences between the paired rates of participants followed a normal distribution. Rates are also presented with paired t-test results and their corresponding Cohen's *d* effect size. Statistical significance is indicated with a *.

The sensitivity (*A'*) and bias (B''_D) results and statistics comparing them are reported in Figure 6. These analyses showed that sensitivity was significantly lower for masking trials than non-masking trials. This means that people had a more difficult time distinguishing between signal and noise when masking was predicted than when it was not. On average, bias measures were positive. This indicates that participants tended to say "Yes" more often than they said "No". People tended to say "Yes" more often for masking trials than for non-masking trials. This difference would have met a 0.05 significance level ($p = 0.008$), however this failed to meet the adjusted level of statistical significance.

Discussion

These results are consistent with our hypotheses. We found that participants made more misses when the test sounds were masked than when they were not. In the non-masking condition, participants had misses only roughly one-third of the time while, in the masking condition, participants made misses on average 48.1% of the time, which is extremely close to 50% (which would be expected by random guessing). Conversely, there was no significant difference in false alarm rates between the two conditions, which happened roughly 30% of the time. Further, participants had reduced sensitivity for masking trials than for non-masking ones. Collectively, these results suggest that participants clearly had more trouble distinguishing between signal and noise in the masking condition than the non-masking one, and that this was predominantly due to the fact that masking makes it more likely that humans will miss alarms. This is an important result because it validates that the psychoacoustics of simultaneous masking that are used in our method are able to accurately predict whether or not masking will contribute to alarm perceivability.

It is slightly concerning that the judgment error rates observed outside of the masking miss condition occurred roughly one-third of the time and were not closer to zero. This is likely due to the fact that the judgment task was difficult and that there are higher perceptual, attentional, and cognitive factors that will influence it. Implications of this are explored in greater depth in the general discussion.

The results on bias did not strictly violate our hypothesis that there would be no statistically significant difference between the masking and non-masking conditions. There does appear to be a trend that people were biased towards saying “Yes” more often in masking trials than in non-masking ones. It is not entirely clear why this occurred. This will be explored in greater depth in the general discussion.

Experiment 2

Experiment 1 provided evidence for the validity of the psychoacoustics of simultaneous masking. However, because this experiment did not include any additional harmonics, it is not clear whether these results would generalize to complete alarm sounds as specified in the international standard (see Table 1). Thus, the second experiment we conducted was designed to see how well the masking of alarm sound primary harmonics impacts the perceivability of more complex alarm sounds that include the requisite additional harmonics dictated by the standard (see Table 1).

Methods

Experiment 2 was almost an identical replication of Experiment 1. It was performed with 28 new nursing student participants (this time with 24 females and 4 males) with the same apparatus, methods, and experimental design. There were two important differences.

First, while the alarm sounds represented the same set of 200 trials from the first experiment, the versions of the sounds used in Experiment 2 were extended to include 4 additional harmonics that played concurrently with the original primary harmonic of the sound. In all cases, these additional harmonics were computed as being 3, 5, 7, and 9 times the frequency of the primary harmonic (whole number multiples are typically used to avoid

dissonance in the complex sound). Each additional harmonic had a volume 15 dB lower than the primary one. These parameters made the alarm sounds compliant with the standard (see Table 1) and were consistent with common recommendations for accomplishing this (Thompson, 2010).

The second difference from Experiment 1 came in the data analysis. While the results of Experiment 2 were evaluated using the same methods as Experiment 1, we also used standard (non-paired) two-tailed t-tests to determine if there were significant differences between comparable measures (M, F , A', and B''_D) between the experiments. This allowed us to assess whether the inclusion of the additional harmonics improve or reduce alarm perceivability in both the presence and absence of masking.

Results

Results and statistics for the miss rate (M) and false alarm rate (F) analyses are shown in Figure 7. These showed that miss rate was significantly higher for the masking condition than for the non-masking one and that there were no statistically significant differences in false alarm rates.

The results of the sensitivity (A') and bias (B''_D) analyses are also shown in Figure 7. These showed that there were significant differences between sensitivity and bias. On average, participants were significantly less sensitive in the masking condition than in the non-masking condition. Conversely, participants had a significantly higher bias (and thus tended to say “Yes”) more often in the masking condition.

The comparison of these SDT statistics to the comparable ones from Experiment 1 (see Figure 8) revealed that there were no significant difference between any of them.

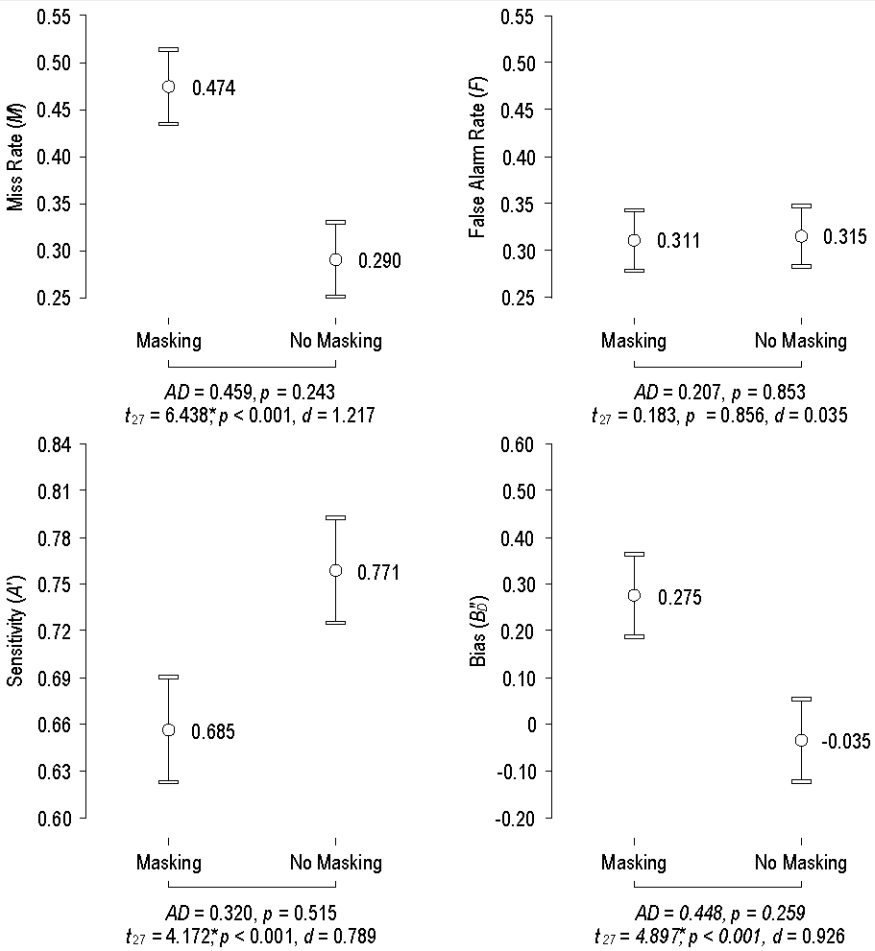

Figure 7: Means (labeled circles) and 95% within-subjects confidence intervals (horizontal bars) for miss rates (***M***), false alarm rates (***F***), sensitivity (***A′***), and bias (***B″D***) for both the masking and non-masking conditions observed during experiment 2. Rates are presented with Anderson Darling statistics that indicate that differences between the paired rates of participants followed a normal distribution. Rates are also presented with paired t-test results and their corresponding Cohen's *d* effect size. Statistical significance is indicated with a *.

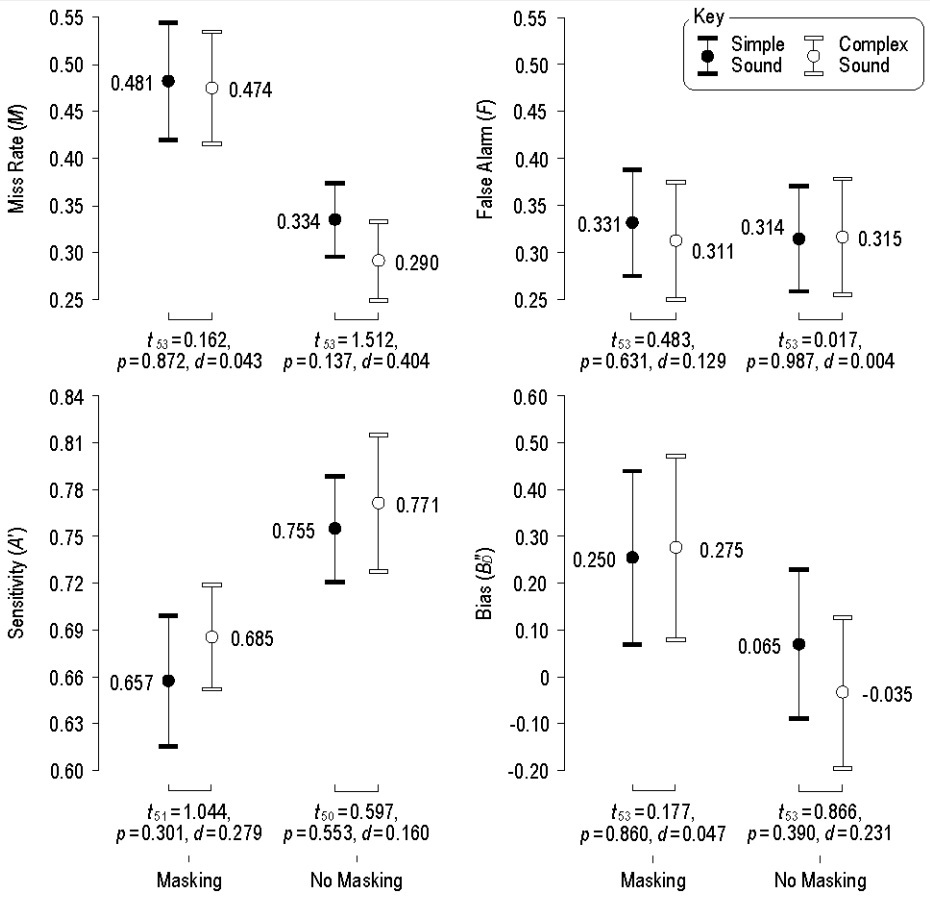

Figure 8: Comparisons of miss rate (*M*), false alarm rate (*F*), sensitivity (*A'*), and bias (B''_D) values measured in experiments 1 and 2 (reported previously in Figures 6 and 7). T-test statistics (reported with their corresponding Cohen's *d* effect size) show that there were no statistically significant differences observed between comparable rates of the two experiments. Note that due to the nature of the comparisons being done, these plots are presented with between-subject confidence intervals which differ from the within-subject confidence intervals presented in Figures 6 and 7.

Discussion

The results for Experiment 2 effectively replicated the results seen for Experiment 1. It produces comparable values between the computed SDT measures and none of the comparable measures difference were statistically significant. These results show that the inclusion of additional harmonics does not impact the overall perceivability of alarms for any of the experimental conditions. Given the comparable rates and sensitivities across the masking and non-masking conditions, this means that the additional harmonics neither counteract the effect of masking nor do they help improve the overall perceivability of the alarms. This is a compelling result that will be discussed further in the next section. It is important to note that because the frequencies of the additional harmonics were obtained by multiplying the primary harmonic by

whole numbers, it is extremely unlikely that any of these harmonics would be masked due to the bark distances this multiplication creates. Thus, this effect is not due to simultaneous masking. It is our hypothesis that the masking of the primary harmonic reduces the saliency of the alarms such that the additional frequencies are not enough for people to identify them. This will need to be investigated more deeply in future research.

The only slight discrepancy in the results between the two experiments was seen in the response bias measures, which did exhibit a significant difference in Experiment 2 (only a non-significant trend was seen in Experiment 1). As with Experiment 1, it is not entirely clear why participants would tend to say "Yes" in the masking condition. This is discussed more in the next section.

General Discussion and Conclusions

This research used human subject experiments to validate that the psychoacoustics of simultaneous masking are able to predict the perceivability of medical alarm sounds. To the best of our knowledge, this is the first research to empirically show that masking is a problem for the current IEC 60601-1-8 alarms. Furthermore, our research showed that the masking effect is strong enough to reduce the audibility of IEC 60601-1-8-compliant alarms by a statistically significant amount, even with the inclusion of the requisite additional harmonics. These are powerful results because they mean that the psychoacoustics of simultaneous masking can be used to make predictions about whether people will be able to hear alarms from the IEC 60601-1-8 international standard and that this can be done with only the primary harmonics of the alarms.

Our results are of import to our method (Bolton, Edworthy, & Boyd, 2018; Bolton, Edworthy, Boyd, Wei, & Zheng, 2018; Bolton et al., 2016; Hasanain et al., 2016, 2014, 2017), which, in turn, has important implications for alarm design and masking in healthcare environments. First, by validating the predictive capabilities of the psychoacoustics that our method uses, we enable the predictive power of our method to be used effectively to design and evaluate medical alarms and its use in our ongoing effort to evaluate and improve the international medical alarm standard (Bolton, Edworthy, & Boyd, 2018). This has the potential to improve the perceivability of medical alarms across the industry and thus improve patient safety and outcomes. Second, although our method can account for additional harmonics, doing so requires more computational time and resources. So pronounced is this, that it has the potential to limit the applicability of the method. Thus, by showing that we only need to account for the primary harmonics in analyses, our results expand the potential usefulness and approachability of our method. This should help enable the use of our method in the analysis of the planned changes to the international standards and by medical device companies designing medical alarms. Third, our results validate the previous findings that have been made using our method. This includes evaluations of the standard's reserved alarm sounds (Bolton, Edworthy, & Boyd, 2018) and standard-compliant alarms used in real telemetry monitoring systems (Bolton, Edworthy, Boyd, Wei, & Zheng, 2018). These analyses found compelling problems with these alarms. Thus, the previous results along with the validation presented in this paper suggest that there could be serious masking problems with the alarms of IEC 60601-1-8. Future work should systematically explore when and how masking can manifest in the standard.

Beyond the masking results, our experiments also provide some troubling data about the standard. In particular, across both experiments, the minimum miss and false alarm rates (even in

the absence of masking) was approximately 30%. This means that even without masking, the alarm sounds prescribed by the standard can be very difficult to distinguish from the others. Although we used a different experimental design, our results are consistent with research by Lacherez, Seah, and Sanderson (2007) who found that alarm sounds from the standard were very difficult to distinguish from each other when they played concurrently. As such, it is clear that changes will need to be made to the sounds of the alarm standard to make them more distinguishable. The work presented in this paper is being conducted concurrently with a number of other coordinated efforts (Edworthy et al., 2018) to address shortcomings in IEC 60601-1-8 and recommend improvements. Thus, results from the work presented in the paper will be used to help improve the general distinguishability of standard alarm sounds.

As with any study, there were some limitations to our experiments. These and future work are discussed in the following sections.

Additional Experimental Considerations

There are factors that limit the realism of our experiment: we only considered single tones from alarm melodies; experiments were conducted in a quiet controlled laboratory (not a realistic environment); and participants were able to give the experiment their undivided attention (something extremely unlikely in a healthcare scenario). All of these factors were intentionally chosen to allow the experiment to isolate the effect of masking and minimize the impact of other limits on human perception, attention, and cognition. However, future work could investigate the true impact masking would have on alarm identification in more realistic contexts. Given the strong impact masking had on detection in the ideal listening conditions in our experiment, we would expect even worse detection performance in more realistic settings. Future work should investigate what proportion of alarm perceivability is attributable to simultaneous masking in realistic medical environments.

Experiment 2 only considered one method for including additional frequencies in alarm sounds. While the parameters for these that were used in our experiment followed common guidelines (Thompson, 2010), it is possible that different parameters could improve alarm distinguishability. In particular, alarms could possibly be made to be more salient by using additional harmonics that are not integer multiples of the primary one, thus creating harmonic dissonance. This should be the subject of future research.

Investigation of Bias

In both experiments, participants had a larger, positive bias in the masking condition than in the non-masking condition (although this difference was only statistically significant in Experiment 2). This means that they tended to say “Yes” in masking trials more often than in the non-masking ones. It is not clear why this occurred. One possibility has to do with the fact that in trials with masking, masking sounds could sometimes sound slightly “warbly” (trilling or quavering). This may be caused by physical interactions (called beating; see Levitin 2006) between the frequencies of the masking and masked sounds. It is possible that this “warbliness” was used by some participants as a cue that the judgment sound was present. This should be investigated in future research.

Additional Alarm Sounds

As part of the larger effort to revise the standard (Edworthy et al., 2018), researchers are designing new alarms that are more complex and harmonically rich than the current melodies of tones, though the melodic patterns will likely remain through legacy support. Thus, while the results presented here will remain topical, the psychoacoustics of simultaneous masking validated in this work are not appropriate for the new alarm sounds. However, there are other masking curves that use different formulations of spreading functions and Δ than those shown in Equations 4 and 5 respectively that can represent the masking effect of more complex sounds (Bosi & Goldberg, 2003; You, 2010). Future work should investigate which of these is most appropriate for the new sounds and use experimental validation (like the one presented here) to assess their predictive power.

Additional Application Domains

The focus of the presented research is exclusively on medical alarms. However, alarms are used to alert humans to problems in many other safety critical domains including aviation (Bliss & Acton, 2003), industrial control rooms (Rothenberg, 2009), and driving (Bliss, 2003). Many of the same problems that impact medical alarms can also manifest in these other areas. In fact, there have been a few instances of design recommendations for avoiding the effects of masking in these industries (Begault, Godfroy, Sandor, & Holden, 2007; Patterson, 1982; Patterson & Mayfield, 1990; Wolfman, Miller, & Volanth, 1996). However, to the best of our knowledge, nobody has investigated whether simultaneous masking does in fact manifests in these environments. Thus, future work should determine whether simultaneous masking is occurring and, if so, how our methods could be used to assess its potential risks.

Key Concepts

- The alarms prescribed by the IEC 60601-1-8 international standard are theoretically susceptible to simultaneous masking.
- This work validates that the psychoacoustics of simultaneous masking can accurately predict the perceivability of standard-compliant medical alarm sounds using two signal detection experiments.
- The experiments showed that the psychoacoustics did accurately predict the perceivability of alarm sounds based on whether their primary harmonics were masked.
- The results further validate that a formal methods model developed in previous work can accurately predict whether humans will hear IEC 60601-1-8-compliant alarms.
- The results will influence methods for detecting masking in medical alarm designs as well as updates to the international standard.

References

Ambikairajah, E., Davis, A., & Wong, W. (1997). Auditory masking and MPEG-1 audio compression. *Electronics & Communication Engineering Journal*, *9*(4), 165–175.

Baumgarte, F., Ferekidis, C., & Fuchs, H. (1995). A nonlinear psychoacoustic model applied to ISO/MPEG layer 3 coder. In *Proceedings of the audio engineering society convention.* New York: Audio Engineering Society.

Begault, D. R., Godfroy, M., Sandor, A., & Holden, K. (2007). Auditory alarm design for NASA CEV applications. In *Proceedings of the 13th international conference on auditory display* (pp. 131–138). Montreal, Canada.

Bliss, J. P. (2003). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology*, *13*(3), 249–268.

Bliss, J. P., & Acton, S. A. (2003). Alarm mistrust in automobiles: How collision alarm reliability affects driving. *Applied Ergonomics*, *34*(6), 499–509.

Bolton, M. L., Edworthy, J. R., & Boyd, A. D. (2018). A formal analysis of masking between reserved alarm sounds of the IEC 60601-1-8 international medical alarm standard. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, pp. 523–527). Los Angeles.

Bolton, M. L., Edworthy, J. R., Boyd, A. D., Wei, J., & Zheng, X. (2018). A computationally efficient formal method for discovering simultaneous masking in medical alarms. *Applied Acoustics*, *141*, 403–415.

Bolton, M. L., Hasanain, B., Boyd, A. D., & Edworthy, J. R. (2016). Using model checking to detect masking in IEC 60601-1-8-compliant alarm configurations. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 636–640). Los Angeles.

Bosi, M., & Goldberg, R. E. (2003). *Introduction to digital audio coding and standards*. New York: Springer.

Brandenburg, K., & Bosi, M. (1997). Overview of MPEG audio: Current and future standards for low bit-rate audio coding. *Journal of the Audio Engineering Society*, *45*(1/2), 4–21.

Brandenburg, K., & Stoll, G. (1994). ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio. *Journal of the Audio Engineering Society*, *42*(10), 780–792.

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model checking*. Cambridge: MIT Press.

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: a comment on baguley (2012). *Behavior Research Methods*, *46*(4), 1149–1151.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, *121*(3), 275–277.

E. Zwicker and R. Feldtkeller. (1967). *Das ohr als nachrichtenempfänger*. Stuttgart: Hirzel Verlag.

ECRI Institute. (2014). Top 10 health technology hazards for 2015. *Health Devices*, *November*. Retrieved from http://www.ecri.org/2015hazards

ECRI Institute. (2018). 2019 top 10 health technology hazards. *Health Devices*, *October*. Retrieved from https://www.ecri.org/Resources/Whitepapers_and_reports/Haz_19.pdf

ECRI Institute, & ISMP. (2009). Connecting remote cardiac monitoring issues with care areas. *Pennsylvania Patient Safety Authority*, *6*(3), 79–83. Retrieved from http://patientsafetyauthority.org/ADVISORIES/AdvisoryLibrary/2009/Sep6(3)/Pages/79.aspx

Edworthy, J. R. (2013). Medical audible alarms: A review. *Journal of the American Medical Informatics Association*, *20*(3), 584–589.

Edworthy, J. R., & Hellier, E. (2005). Fewer but better auditory alarms will improve patient safety. *Quality and Safety in Health Care*, *14*(3), 212–215.

Edworthy, J. R., & Hellier, E. (2006). Alarms and human behaviour: Implications for medical alarms. *British Journal of Anaesthesia*, *97*(1), 12–17.

Edworthy, J. R., McNeer, R. R., Bennett, C. L., Dudaryk, R., McDougall, S. J., Schlesinger, J. J., … others (2018). Getting better hospital alarm sounds into a global standard. *Ergonomics in Design*, *26*(4), 4–13.

Edworthy, J. R., & Meredith, C. S. (1994). Cognitive psychology and the design of alarm sounds. *Medical Engineering & Physics*, *16*(6), 445–449.

Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: Facts and models* (Vol. 22). Springer.

Green, D. M. (1967). Additivity of masking. *The Journal of the Acoustical Society of America*, *41*(6), 1517–1525.

Hasanain, B., Boyd, A., & Bolton, M. (2016). Using model checking to detect simultaneous masking in medical alarms. *IEEE Transactions on Human-Machine Systems*, *46*(2), 174–185.

Hasanain, B., Boyd, A., & Bolton, M. L. (2014). An approach to model checking the perceptual interactions of medical alarms. In *Proceedings of the 2014 international annual meeting of the human factors and ergonomics society* (pp. 822–826). Santa Monica: HFES.

Hasanain, B., Boyd, A. D., Edworthy, J. R., & Bolton, M. L. (2017). A formal approach to discovering simultaneous additive masking between auditory medical alarms. *Applied Ergonomics*, *58*, 500–514.

Humes, L. E., & Jesteadt, W. (1989). Models of the additivity of masking. *The Journal of the Acoustical Society of America*, *85*(3), 1285–1294.

IEC 60601-1-8:2006+AMD1:2012. (2012). *Medical electrical equipment - part 1-8: General requirements for basic safety and essential performance – collateral standard: General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems*. Geneva: International Electrotechnical Commission.

Konkani, A., Oakley, B., & Bauld, T. J. (2012). Reducing hospital noise: A review of medical device alarm management. *Biomedical Instrumentation & Technology*, *46*(6), 478–487.

Lacherez, P., Seah, E., & Sanderson, P. (2007). Overlapping melodic alarms are almost indiscriminable. *Human Factors*, *49*(4), 637–645.

Levitin, D. J. (2006). *This is your brain on music: The science of a human obsession*. Penguin.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401-413.

McNicol, D. (2005). *A primer of signal detection theory*. Mahwah: Lawrence Erlbaum.

Meredith, C., & Edworthy, J. R. (1995). Are there too many alarms in the intensive care unit? An overview of the problems. *Journal of Advanced Nursing*, *21*(1), 15–20.

Momtahan, K., Hetu, R., & Tansley, B. (1993). Audibility and identification of auditory alarms in the operating room and intensive care unit. *Ergonomics*, *36*(10), 1159–1176.

Patterson, R. D. (1982). *Guidelines for auditory warning systems on civil aircraft*. London.

Patterson, R. D., & Mayfield, T. F. (1990). Auditory warning sounds in the work environment. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *327*(1241), 485–492.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1-12), 125–126.

Rothenberg, D. H. (2009). *Alarm management for process control: A best-practice guide for design, implementation, and use of industrial alarm systems*. New York: Momentum Press.

Schroeder, M. R., Atal, B. S., & Hall, J. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, *66*, 1647–1652.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors*, *39*(1), 14–29.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Stead, W. W., & Lin, H. S. (Eds.). (2009). *Computational technology for effective health care: Immediate steps and strategic directions*. Atlanta: National Academies Press.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, *1*(2), 155–182.

The Joint Commission. (2013a, April). Medical device alarm safety in hospitals. *Sentinel Even Alert*, *50*.

The Joint Commission. (2013b, July). NPSG.06.01.01: Improve the safety of clinical alarm systems. *Joint Commission Perspectives*, *33*.

Thompson, C. (2010). *ISO/IEC 60601-1-8, Patterson and other alarms in medical equipment sample alarm sounds - sirens, buzzers and other sounds.* (`http://www.anaesthesia.med.usyd.edu.au/resources/alarms/`)

Toor, O., Ryan, T., & Richard, M. (2008). Auditory masking potential of common operating room sounds: A psychoacoustic analysis. In *Anesthesiology* (Vol. 109, p. A1207). Park Ridge: American Society of Anesthesiologists.

Vasseur, D. A., & Yodzis, P. (2004). The color of environmental noise. *Ecology*, *85*(4), 1146–1152.

Vockley, M. (2014). *Clinical alarm management compendium*. Arlington: AAMI Foundation.

Wolfman, G. J., Miller, D. L., & Volanth, A. J. (1996). An application of auditory alarm research in the design of warning sounds for an integrated tower air traffic control computer system. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 40, pp. 1002–1006).

You, Y. (2010). *Audio coding: Theory and applications*. Springer Science & Business Media.

Biographies

Matthew L. Bolton is an Associate Professor of Industrial and System Engineering at the University at Buffalo, the State University of New York. He received the Ph.D. in Systems Engineering in 2010 from the University of Virginia, Charlottesville, USA.

Xi Zheng is a Ph.D. Student in Industrial and Systems Engineering at the University at Buffalo, the State University of New York. She received the B.S. in electronic commerce in 2011 from Southwest University, Chongqing, China.

Meng Li is a Human Factors Engineer and UX Designer Intern at Medtronic. He received the Ph.D. in Industrial and Systems Engineering in 2018 from the University at Buffalo, the State University of New York, Buffalo, USA.

Judy Reed Edworthy is the Director of the Cognition Institute and a Professor of Applied Psychology at the University of Plymouth. She received the Ph.D. in Experimental Psychology in 1984 from the University of Warwick, UK

Andrew D. Boyd is an Associate Professor of Biomedical and Health Information Sciences at the University of Illinois Chicago. He received the M.D. in 2002 from the University of Texas Southwestern Medical School, Dallas.